HPC: Project Report

2017-19841 최창민

병렬화 방식

GPU 커널 구현 및 Stream

저는 최대한 대부분의 작업을 GPU에서 진행하려고 했습니다. 그래서 대부분의 연산들을 gpu로 진행했습니다. 또 거의 모든 연산을 Fusion하였습니다. 대부분은 거의 간단하게 구현되었는데, softmax를 위해서 reduce_sum하는 부분이 약간 어려웠습니다. reduce sum을 하기 위해서 한 커널을 여러번 호출했고, 커널 안에서는 stride 하면서 reduce sum을 했습니다. 또 다른 것은 거의 다 GPU에서 할 수 있었지만 argmax인 top_one은 gpu상에서 구현하기 힘들어 CPU에서 작업했습니다. 그런데 이런식으로 하면 계속 CPU에서 blocking이 걸려서 pthread를 이용하여 한 GPU에서 여러 Stream을 돌릴 수 있게 했습니다.

GPU / MPI 병렬화 대상

기본적으로 저는 한 문장은 한 노드 안의 한 GPU의 한 Stream이 담당하도록 했습니다. 그래서 문장들을 Node, GPU에 다 분배한 다음 다시 Gather했습니다. 분배에서는 Async MPI Communication을 쓴 다음 연산하기 직전에 Waitall로 기다렸습니다. GPU의 경우에는 정확하게는 따로 분배하진 않고 index를 다르게 두어 접근할 수 있게 했습니다.

제가 측정한 성능